

**KLASIFIKASI DOKUMEN BERITA BAHASA INDONESIA  
MENGUNAKAN METODE *LATENT DIRICHLET ALLOCATION*  
(LDA) DAN *WORD2VEC***



**SKRIPSI**

**Disusun Sebagai Salah Satu Syarat  
untuk Memperoleh Gelar Sarjana Komputer  
pada Departemen Ilmu Komputer/ Informatika**

**Disusun oleh :  
SAMUEL ADI PRASETYO  
24010313140099**

**DEPARTEMEN ILMU KOMPUTER/ INFORMATIKA  
FAKULTAS SAINS DAN MATEMATIKA  
UNIVERSITAS DIPONEGORO  
2018**

## HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Saya yang bertanda tangan di bawah ini :

Nama : Samuel Adi Prasetyo  
NIM : 24010313140099  
Judul : Klasifikasi Dokumen Berita Bahasa Indonesia Menggunakan Metode  
*Latent Dirichlet Allocation (LDA) dan Word2Vec*

Dengan ini saya menyatakan bahwa dalam tugas akhir ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan di dalam daftar pustaka.

Semarang, 4 Juni 2018



Samuel Adi Prasetyo

NIM. 24010313140099

## HALAMAN PENGESAHAN

Judul : Klasifikasi Dokumen Berita Bahasa Indonesia Menggunakan Metode  
*Latent Dirichlet Allocation (LDA)* dan *Word2Vec*  
Nama : Samuel Adi Prasetyo  
NIM : 24010313140099

Telah diujikan pada sidang tugas akhir pada tanggal 22 Mei 2018 dan dinyatakan lulus pada tanggal 22 Mei 2018.

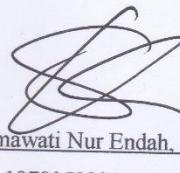
Semarang, 4 Juni 2018

Mengetahui,  
Ketua Departemen Ilmu Komputer/ Informatika



Dr. Retno Kusumaningrum, S.Si, M.Kom  
NIP. 198104202005012001

Panitia Penguji Tugas Akhir  
Ketua,



Sukmawati Nur Endah, S.Si, M.Kom  
NIP. 197805022005012002

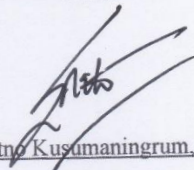
## HALAMAN PENGESAHAN

Judul : Klasifikasi Dokumen Berita Bahasa Indonesia Menggunakan Metode  
*Latent Dirichlet Allocation (LDA)* dan *Word2Vec*  
Nama : Samuel Adi Prasetyo  
NIM : 24010313140099

Telah diujikan pada sidang tugas akhir pada tanggal 22 Mei 2018.

Semarang, 4 Juni 2018

Dosen Pembimbing



Dr. Retno Kusumaningrum, S.Si, M.Kom

NIP. 198104202005012001

## ABSTRAK

Perkembangan yang pesat dalam informasi digital telah menyebabkan semakin meningkat pula volume informasi yang berbentuk teks seperti dokumen berita. Dokumen berita yang muncul diunggah di internet sangatlah banyak dalam rentang waktu yang cepat. Oleh karena itu diperlukan adanya pengorganisasian dokumen berita. Salah satu cara yang dapat dilakukan dengan cepat dan dapat dipahami oleh para penerima informasi adalah dengan melakukan klasifikasi dokumen berita berdasarkan topiknya. Penelitian yang diusulkan yaitu penerapan klasifikasi dokumen untuk berita Bahasa Indonesia menggunakan metode *Latent Dirichlet Allocation* (LDA) yang akan digabungkan dengan metode *word embedding Word2Vec* dan juga *k-means clustering* sebagai metode pembantu melakukan klusterisasi vektor kata. Dokumen berita Bahasa Indonesia akan diklasifikasikan ke dalam lima topik yaitu olahraga, teknologi, ekonomi, politik, dan sosial dimana kelima kategori tersebut merupakan kategori berita utama yang sering diakses oleh pengguna. Hasil penelitian dengan jumlah data pelatihan sebanyak 1000 berita (200 berita per kategori) menunjukkan bahwa metode gabungan LDA dan *Word2Vec* sudah cukup baik dalam melakukan klasifikasi dengan nilai akurasi tertinggi sebesar 73,4%. Meski demikian akurasi lebih baik didapatkan oleh metode LDA murni tanpa *Word2Vec* dengan nilai akurasi sebesar 87,5% sehingga memiliki selisih akurasi sebesar 14,1%. Kedua perbandingan metode tersebut sama-sama diperoleh pada kombinasi parameter  $\alpha$  0,1;  $\beta$  0,01; dan jumlah topik sebanyak 300 topik.

**Kata Kunci:** *Text Mining*, Klasifikasi Berita Bahasa Indonesia, *Latent Dirichlet Allocation*, *Word2Vec*, *K-means Clustering*

## ***ABSTRACT***

The rapid development of digital information has led to an over-increasing volume of textual information such as news documents. News documents that uploaded to internet are very massive in the short of time. Therefore it is necessary to organize news documents. One way that can be done quickly and comprehensible by the recipients of information is to classify news documents based on their topic. The proposed research is the application of document classification for Indonesian news using Latent Dirichlet Allocation (LDA) method which will be combined with Word2Vec word embedding method and also k-means clustering as support method to cluster word's vector. Indonesian news documents will be classified into five topics: sports, technology, economics, politics, and social where the five categories are major news categories that are often accessed by users. Result of the research with amount of training data of 1000 news (200 news per category) showed that the combined method of LDA and Word2Vec was good enough in classification with the highest accuracy value of 73,4%. However better accuracy is obtained by the pure LDA method without Word2Vec with an accuracy value of 87,5%, thus having an accuracy difference of 14,1%. Both comparison methods were similarly obtained on the parameter combination of alpha 0,1; beta 0,01; and 300 number of topics.

**Keywords:** Text Mining, Indonesian News Document Classification, Latent Dirichlet Allocation, Word2Vec, K-means Clustering

## KATA PENGANTAR

Puji syukur kepada Tuhan Yang Maha Esa atas berkat dan penyertaan-Nya sehingga penulis dapat menyelesaikan skripsi dengan judul “Klasifikasi Dokumen Berita Bahasa Indonesia Menggunakan Metode *Latent Dirichlet Allocation* (LDA) dan *Word2Vec*” dengan baik dan lancar.

Tugas Akhir ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana Strata Satu (S1) pada Departemen Ilmu Komputer/ Informatika Fakultas Sains dan Matematika Universitas Diponegoro Semarang.

Pada penyusunan Tugas Akhir ini, penulis banyak mendapat bimbingan, arahan dan bantuan dari berbagai pihak. Oleh karena itu, melalui kesempatan kali ini penulis mengucapkan rasa hormat dan terima kasih kepada:

1. Ibu Dr. Retno Kusumaningrum, S.Si, M.Kom, selaku Ketua Departemen Ilmu Komputer/ Informatika FSM Universitas Diponegoro Semarang, sekaligus menjadi Dosen Pembimbing yang telah membantu dalam proses bimbingan hingga selesainya Laporan Tugas Akhir ini.
2. Bapak Helmie Arif Wibawa, S.Si, M.Cs, selaku Koordinator Tugas Akhir Departemen Ilmu Komputer/Informatika FSM Universitas Diponegoro Semarang.
3. Bapak Tri Agus Rahmanto dan Ibu Maryati, selaku orang tua yang terus memberikan doa dan dukungan baik secara moril dan materiil dalam menyelesaikan Tugas Akhir ini.
4. Ditya Hanggara, Julius Evans, dan Widy Ageng, sebagai sahabat yang saling mendukung selama menempuh studi di Informatika ini.
5. Semua pihak yang telah membantu kelancaran dalam penyusunan Tugas Akhir, yang tidak dapat disebutkan satu persatu.

Pada pembuatan Laporan Tugas Akhir ini masih memiliki banyak kekurangan baik dari segi materi ataupun dari segi penyajiannya karena keterbatasan kemampuan dan pengetahuan dari penulis. Oleh karena itu, kritik dan saran sangat diharapkan. Terima kasih, dan semoga Laporan Tugas Akhir ini bermanfaat bagi semua pihak.

Semarang, 22 Mei 2018

Samuel Adi Prasetyo

## DAFTAR ISI

<b>HALAMAN PERNYATAAN KEASLIAN SKRIPSI .....</b>	<b>ii</b>
<b>HALAMAN PENGESAHAN .....</b>	<b>iii</b>
<b>HALAMAN PENGESAHAN .....</b>	<b>iv</b>
<b>ABSTRAK.....</b>	<b>v</b>
<b>ABSTRACT.....</b>	<b>vi</b>
<b>KATA PENGANTAR .....</b>	<b>vii</b>
<b>DAFTAR ISI .....</b>	<b>viii</b>
<b>DAFTAR GAMBAR .....</b>	<b>x</b>
<b>DAFTAR TABEL.....</b>	<b>xii</b>
<b>DAFTAR LAMPIRAN .....</b>	<b>xiv</b>
<b>BAB I PENDAHULUAN .....</b>	<b>1</b>
1.1. Latar Belakang.....	1
1.2. Rumusan Masalah.....	3
1.3. Tujuan dan Manfaat .....	3
1.4. Ruang Lingkup .....	4
1.5. Sistematika Penulisan .....	4
<b>BAB II TINJAUAN PUSTAKA .....</b>	<b>6</b>
2.1. Klasifikasi Teks Berita Bahasa Indonesia .....	6
2.2. <i>Preprocessing</i> .....	7
2.2.1. <i>Case Folding</i> .....	7
2.2.2. <i>Tokenization</i> .....	8
2.2.3. <i>Filtering</i> .....	8
2.2.4. <i>Stemming</i> .....	9
2.3. <i>Word2Vec Model</i> .....	17
2.4. <i>K-Fold Cross Validation</i> .....	21
2.5. <i>K-Means Clustering</i> .....	22
2.6. <i>Latent Dirichlet Allocation (LDA)</i> .....	23
2.6.1. Proses Pelatihan <i>Latent Dirichlet Allocation (LDA)</i> .....	27
2.6.2. Proses Pengujian <i>Latent Dirichlet Allocation (LDA)</i> .....	28
2.7. <i>Kullback Leibler Divergence</i> .....	29
2.8. Evaluasi.....	30



<b>BAB III METODOLOGI PENELITIAN.....</b>	<b>32</b>
3.1. Pengumpulan Data.....	33
3.2. <i>Preprocessing</i> .....	34
3.2.1. <i>Case Folding</i> .....	34
3.2.2. <i>Tokenization</i> .....	35
3.2.3. <i>Filtering</i> .....	37
3.2.4. <i>Stemming</i> .....	39
3.3. <i>Word2Vec &amp; K-Means Clustering</i> .....	45
3.4. Pembentukan Model <i>Latent Dirichlet Allocation (LDA)</i> .....	46
3.4.1. Pembagian Data dan Pengindeksan .....	47
3.4.2. Pelatihan .....	48
3.4.3. Pengujian .....	54
3.5. Evaluasi.....	57
3.6. Pembentukan Aplikasi Klasifikasi Dokumen Berita Bahasa Indonesia .....	59
3.7. Analisis dan Desain Aplikasi .....	60
3.7.1. Analisis Aplikasi .....	60
3.7.2. Perancangan Aplikasi .....	61
<b>BAB IV HASIL DAN ANALISA .....</b>	<b>70</b>
4.1. Hasil Pengembangan Aplikasi .....	70
4.1.1. Lingkungan Implementasi.....	70
4.1.2. Implementasi Antarmuka .....	71
4.2. Skenario Pengujian Aplikasi.....	76
4.2.1. Skenario Pengujian Fungsional Aplikasi .....	76
4.2.1. Skenario Pengujian Kinerja Aplikasi .....	78
4.3. Hasil dan Analisa Aplikasi .....	79
4.3.1. Hasil dan Analisa Pengujian Fungsional Aplikasi .....	80
4.3.2. Hasil dan Analisa Pengujian Kinerja Aplikasi .....	80
<b>BAB V PENUTUP .....</b>	<b>87</b>
5.1. Kesimpulan .....	87
5.2. Saran .....	88
<b>DAFTAR PUSTAKA .....</b>	<b>89</b>
<b>LAMPIRAN-LAMPIRAN.....</b>	<b>92</b>

## DAFTAR GAMBAR

Gambar 2. 1 Arsitektur <i>Word2vec Model</i> .....	18
Gambar 2. 2 Alur Pembentukan Informasi Dengan <i>Topic Models</i> (Liu, 2013) .....	23
Gambar 2. 3 <i>Graphical Model</i> Dari PLSA Dan LDA .....	24
Gambar 2. 4 Model LDA <i>Generative</i> dan <i>Inference</i> .....	25
Gambar 2. 5 Visualisasi LDA Sebagai Inferensi (Kusumaningrum, et al., 2014) .....	27
Gambar 2. 6 <i>Confusion Matrix</i> Dua Kelas .....	30
Gambar 3. 1 Gambaran Umum Penelitian .....	32
Gambar 3. 2 <i>Flowchart Preprocessing</i> .....	34
Gambar 3. 3 <i>Flowchart</i> Tokenisasi .....	37
Gambar 3. 4 <i>Flowchart Filtering</i> .....	38
Gambar 3. 5 <i>Flowchart Stemming</i> .....	40
Gambar 3. 6 Sub-Proses <i>Stemming</i> Sastrawi (Bashri, 2017).....	40
Gambar 3. 7 <i>Flowchart Stemming PluralWord</i> (Bashri, 2017).....	41
Gambar 3. 8 <i>Flowchart Stemming SingularWord</i> (Bashri, 2017) .....	42
Gambar 3. 9 <i>Flowchart LoopPengembalianAkhiran</i> (Bashri, 2017) .....	44
Gambar 3. 10 Contoh Vektor Output <i>Word2vec</i> .....	45
Gambar 3. 11 <i>Flowchart Clustering BoW</i> .....	46
Gambar 3. 12 <i>Flowchart</i> Pembentukan Model LDA .....	47
Gambar 3. 13 <i>Flowchart</i> Pelatihan LDA .....	49
Gambar 3. 14 Proses LDA <i>Collapsed Gibbs Sampling</i> .....	50
Gambar 3. 15 <i>Flowchart</i> Pengujian.....	57
Gambar 3. 16 DCD KDBBI .....	62
Gambar 3. 17 DCD Level 1 KDBBI .....	63
Gambar 3. 18 DFD Level 2 KDBBI.....	64
Gambar 3. 19 Desain Antarmuka Halaman Beranda .....	65
Gambar 3. 20 Desain Antarmuka Halaman Data Berita .....	66
Gambar 3. 21 Desain Antarmuka Halaman Data Preproses.....	66
Gambar 3. 22 Desain Antarmuka Halaman Analisa.....	67
Gambar 3. 23 Desain Antarmuka Halaman <i>Input Parameter</i> .....	68
Gambar 3. 24 Desain Antarmuka Halaman Analisa Terinci .....	68
Gambar 3. 25 Desain Antarmuka Halaman <i>Input</i> Berita.....	69

Gambar 3. 26 Desain Antarmuka Halaman Hasil Klasifikasi .....	69
Gambar 4. 1 Implementasi Halaman Beranda .....	71
Gambar 4. 2 Implementasi Halaman Data Berita .....	72
Gambar 4. 3 Implementasi Halaman <i>Bag of Words</i> .....	72
Gambar 4. 4 Implementasi Halaman Vektor <i>Word2Vec</i> .....	73
Gambar 4. 5 Implementasi Halaman <i>Clustering</i> .....	73
Gambar 4. 6 Implementasi Halaman Analisa .....	74
Gambar 4. 7 Implementasi Halaman <i>Input Parameter</i> .....	74
Gambar 4. 8 Implementasi Halaman Analisa Terinci .....	75
Gambar 4. 9 Implementasi Halaman <i>Input Berita</i> .....	75
Gambar 4. 10 Implementasi Halaman Hasil Klasifikasi .....	76
Gambar 4. 11 Rincian Jumlah Data Berita .....	78
Gambar 4. 12 Grafik Akurasi Kombinasi Keseluruhan .....	81
Gambar 4. 13 Grafik Akurasi Kombinasi LDA Terbaik .....	82
Gambar 4. 14 Grafik Akurasi Kombinasi <i>Word2Vec</i> + LDA Terbaik .....	83
Gambar 4. 15 Grafik Perbandingan Akurasi Jumlah Topik .....	84
Gambar 4. 16 Grafik Perbandingan Akurasi Nilai <i>Alpha</i> .....	85
Gambar 4. 17 Grafik Perbandingan Akurasi Nilai <i>Beta</i> .....	86

## DAFTAR TABEL

Tabel 2. 1 <i>State Of The Art</i> .....	6
Tabel 2. 2 Contoh Proses Tokenisasi.....	8
Tabel 2. 3 Contoh Proses <i>Filtering</i> .....	8
Tabel 2. 4 Kombinasi Awalan dan Akhiran yang Dilarang .....	10
Tabel 2. 5 Aturan Pemenggalan Awalan Algoritma <i>Stemming</i> Nazief Adriani.....	12
Tabel 2. 6 Aturan Pemenggalan Awalan Algoritma <i>Stemming Confix Stripping</i> .....	13
Tabel 2. 7 Aturan <i>RulePredence</i> .....	14
Tabel 2. 8 Aturan Pemenggalan Awalan Algoritma <i>Stemming</i> ECS .....	15
Tabel 2. 9 Aturan Pemenggalan Awalan Algoritma <i>Stemming Modified</i> ECS.....	16
Tabel 2. 10 Aturan Pemenggalan Awalan <i>Stemmer</i> Sastrawi .....	17
Tabel 2. 11 Keterangan Notasi dan Definisi Persamaan .....	28
Tabel 2. 12 Model <i>Confusion Matrix</i> Klasifikasi Dokumen Berita Bahasa Indonesia Menggunakan Metode LDA dan <i>Word2Vec</i> .....	30
Tabel 2. 13 Keterangan Notasi dan Definisi Persamaan <i>Accuracy</i> .....	31
Tabel 3. 1 Dataset Berita Sebelum <i>Case Folding</i> .....	34
Tabel 3. 2 Dataset Berita Setelah <i>Case Folding</i> .....	35
Tabel 3. 3 Dataset Berita Sebelum Tokenisasi .....	35
Tabel 3. 4 Dataset Berita Setelah Tokenisasi .....	36
Tabel 3. 5 Contoh <i>Stopwords</i> Bahasa Indonesia .....	39
Tabel 3. 6 Contoh <i>Stemming SingularWord</i> .....	44
Tabel 3. 7 Kombinasi Parameter Eksperimen .....	51
Tabel 3. 8 Contoh Kosakata .....	52
Tabel 3. 9 Nilai <i>Topic Proportion</i> (PZD) .....	53
Tabel 3. 10 Hasil Perhitungan Nilai PZC .....	54
Tabel 3. 11 Contoh <i>Topic Proportion</i> (PZD) Uji .....	55
Tabel 3. 12 Contoh <i>Topic Proportion</i> per Kelas (PZC) Latih.....	55
Tabel 3. 13 Contoh Dataset Asli dan Dataset Prediksi.....	58
Tabel 3. 14 Contoh Perhitungan <i>Confusion Matrix</i> .....	59
Tabel 3. 15 Kebutuhan Fungsional.....	61
Tabel 3. 16 Kebutuhan Non-Fungsional .....	61
Tabel 4. 1 Skenario Pengujian Fungsional Aplikasi .....	77

Tabel 4. 2 <i>Confusion Matrix</i> Kombinasi LDA-K21 .....	82
Tabel 4. 3 <i>Confusion Matrix</i> Kombinasi W2V-K21 .....	83

## DAFTAR LAMPIRAN

Lampiran 1 Perhitungan <i>Word2Vec</i> .....	93
Lampiran 2 Perhitungan <i>Clustering</i> .....	99
Lampiran 3 Perhitungan <i>Latent Dirichlet Allocation (LDA) Collapsed Gibbs Sampling</i>	104
Lampiran 4 Hasil Pengujian Fungsional Aplikasi .....	126
Lampiran 5 Tabel Hasil Akurasi Menggunakan Metode <i>Latent Dirichlet Allocation (LDA)</i> .....	129
Lampiran 6 Tabel Hasil Akurasi Menggunakan Metode <i>Word2Vec</i> Dan <i>Latent Dirichlet</i> <i>Allocation (LDA)</i> .....	131

# **BAB I**

## **PENDAHULUAN**

Bab pendahuluan membahas mengenai latar belakang, rumusan masalah, tujuan dan manfaat, ruang lingkup, serta sistematika penulisan pelaksanaan Tugas Akhir Klasifikasi Dokumen Berita Bahasa Indonesia Menggunakan Metode *Latent Dirichlet Allocation* (LDA) Dan *Word2Vec*.

### **1.1. Latar Belakang**

Teknologi informasi telah berkembang dengan sangat pesat di dunia, semua dimensi dalam kehidupan sekarang tidak dapat terlepas dari penggunaan teknologi. Perkembangan yang pesat dalam informasi digital telah menyebabkan semakin meningkat pula volume informasi yang berbentuk teks seperti dokumen berita. Dokumen berita yang muncul diunggah di internet sangatlah banyak dalam rentang waktu yang cepat terlebih dengan adanya situs-situs berita *online* yang dikelola baik secara profesional maupun secara amatir. Pada umumnya situs-situs berita yang belum dikelola secara profesional tidak melakukan pengorganisasian berita sesuai dengan topik pembahasan berita tersebut dan terdapat ketidaksesuaian isi berita dengan judulnya. Hal tersebut sering disebut dengan *yellow journalism*, yaitu jurnalisme pemburukan makna melalui gambar dan judul yang bombastis dengan tujuan untuk mencoba mencari pembaca dalam jumlah yang besar. Konsumen dari jurnalisme ini adalah kalangan menengah ke bawah yang tingkat pendidikannya tidak tinggi (Nurudin, 2009).

Berdasarkan permasalahan tersebut diperlukan adanya pengorganisasian dokumen berita. Salah satu cara yang dapat dilakukan dengan cepat dan dapat dipahami oleh para penerima informasi adalah dengan melakukan klasifikasi dokumen berita berdasarkan topiknya. Klasifikasi merupakan salah satu metode dalam *text mining* yang bertujuan untuk mendefinisikan kelas dari sebuah objek yang belum diketahui kelasnya. Dengan melakukan klasifikasi dokumen berita berdasarkan topiknya maka penerima informasi dapat mengetahui topik yang diberitakan dalam dokumen tersebut dan dapat menemukan berita lain yang terkait serta meminimalisir *yellow journalism* pada saat mengakses situs portal berita amatir. Selain itu dengan memanfaatkan klasifikasi dokumen berita, baik penulis

berita maupun pembaca berita dapat mengetahui distribusi topik berita pada situs yang dikelola/diakses.

Penelitian mengenai klasifikasi dokumen berita telah dilakukan oleh Widodo, dkk (2016) dengan menggunakan metode multi-label berbasis *domain specific ontology* memberikan hasil akurasi masing-masing sebesar 93,85% untuk kategori olahraga dan 96,32% untuk kategori teknologi. Selain itu, pada penelitian tersebut juga melakukan pengukuran nilai *f-measure* dengan hasil masing-masing sebesar 74,74% untuk kategori olahraga dan 78,96% untuk kategori teknologi.

Penelitian lainnya dilakukan oleh Ariadi dan Fithriasari (2015) membandingkan dua metode yaitu *Naive Bayes* dan *Support Vector Machine* (SVM) menggunakan *stemming Confix Stripping Stemmer*. Penelitian tersebut memberikan hasil akurasi, *precision*, *recall*, dan *f-measure* sebesar 82,2%, 83,9%, 82,2%, dan 82,4% untuk metode *Naive Bayes*, sedangkan untuk metode SVM memberikan hasil akurasi, *precision*, *recall*, dan *f-measure* sebesar 88,1%, 89,1%, 88,1%, dan 88,3%.

Beberapa metode klasifikasi teks yang telah disebutkan sebelumnya masih memiliki kelemahan-kelemahan. Pada metode ontologi hasil nilai akurasi sangat bergantung pada pendefinisian konsep atau kelas yang dilakukan oleh peneliti. Metode *Naive Bayes* sudah banyak diimplementasikan, menggunakan variabel bebas, hasil klasifikasi bergantung pada fitur yang digunakan, dan tidak berlaku jika probabilitas kondisionalnya adalah nol. Sedangkan pada metode SVM banyak teks tidak dapat diklasifikasikan dengan benar karena adanya masalah *sparseness* data yang disebabkan oleh karakteristik dimensi yang tinggi, masih kaku, dan kinerja tergantung pada pemilihan fungsi kernel (Kusumaningrum et al., 2016). *Sparseness* data merupakan keadaan dimana terdapat kumpulan data yang mengandung nilai mendekati nol lebih dominan.

Oleh karena itu, diperlukan metode reduksi pada dimensi data yang besar dan implementasi klasifikasi dokumen berita Bahasa Indonesia dengan konsep *topic modelling*, yang di antaranya *Latent Semantic Analysis* (LSA), *Probabilistic Latent Semantic Analysis* (PLSA) dan *Latent Dirichlet Allocation* (LDA). LSA mempunyai kekurangan dalam mengolah data dengan jumlah yang besar. PLSA merupakan pembaharuan dari LSA, PLSA dapat mencari topik yang tersembunyi pada korpus (Hofmann, 1999). *Latent Dirichlet Allocation* (LDA) menggunakan model hirarki sehingga lebih stabil dan dapat mengolah data dalam jumlah besar (Liu, 2013).



Metode LDA mengasumsikan bahwa pada satu dokumen terdapat lebih dari satu topik, yang masing-masing merupakan distribusi melalui kosakata. LDA dapat digunakan sebagai *feature selection* dengan tujuan menemukan struktur semantik yang mendasarinya (Li *et al.*, 2011). Dalam pengklasifikasian dokumen, LDA dapat menghasilkan hasil yang cukup baik (Blei, 2012). Penerapan klasifikasi dokumen berita bahasa Indonesia dengan metode LDA oleh Kusumaningrum *et al.*, (2016) menghasilkan akurasi sebesar 70% dengan jumlah dataset pelatihan hanya sebanyak 100 berita.

Dalam banyak kasus, klasifikasi teks dapat sulit diukur karena seiring bertambahnya kata baru, jumlah pelatihan yang diperlukan juga meningkat. Selain itu, dengan jumlah data yang banyak, akan menjadi semakin sulit untuk mengumpulkan contoh teks berlabel untuk setiap kelas. Salah satu penyelesaian yang dapat diusulkan untuk permasalahan ini yaitu dengan penggunaan *Word2Vec* dalam memproses data teks yang tidak terstruktur. *Word2Vec* mengambil setiap kata dalam kosakata dengan merepresentasikannya ke dalam bentuk vektor yang dapat ditambahkan, dikurangkan, dan dimanipulasi. Algoritma ini dapat mengefisiensikan implementasi dari arsitektur untuk komputasi representasi kata menjadi vektor. Nilai vektor dari kata tersebut dapat digunakan untuk memetakan kata dalam kosakata yang saling berkaitan (Tamir, 2016). Model *Word2Vec* merupakan salah satu aplikasi baru pada *machine learning* yang berfokus pada pemrosesan teks yang menarik banyak perhatian sebagai objek penelitian (Meyer, 2015), untuk itu penelitian ini mencoba melakukan penggabungan model *Word2Vec* dengan LDA sebagai metode klasifikasi dokumen berita Bahasa Indonesia.

## **1.2. Rumusan Masalah**

Berdasarkan latar belakang yang telah dijelaskan, maka dapat dirumuskan permasalahan yaitu bagaimana menerapkan metode *Latent Dirichlet Allocation* (LDA) dan *Word2Vec* untuk melakukan klasifikasi dokumen berita Bahasa Indonesia.

## **1.3. Tujuan dan Manfaat**

Tujuan dari penelitian ini adalah membandingkan penerapan metode LDA menggunakan *Word2Vec* dan tanpa *Word2Vec* pada klasifikasi dokumen berita

Bahasa Indonesia. Hasil dari penelitian Tugas Akhir ini diharapkan dapat memberikan kontribusi terhadap penelitian mengenai klasifikasi teks berita bahasa Indonesia dengan kombinasi metode LDA dan *Word2Vec*.

#### 1.4. Ruang Lingkup

Ruang lingkup dari Klasifikasi Teks Berita Bahasa Indonesia menggunakan Metode *Latent Dirichlet Allocation* (LDA) dengan *Word2Vec* adalah sebagai berikut:

1. *Dataset* berita pelatihan dari artikel berita bahasa Indonesia disimpan ke dalam format csv secara manual.
2. Berita yang dibutuhkan untuk pelatihan sebanyak 1000 data dengan masing-masing kategori sebanyak 200 yang diperoleh dari portal berita *online* Detik dan Kompas dari bulan Juli 2017 sampai Oktober 2017.
3. Penentuan kategori dokumen berita berupa olahraga, teknologi, ekonomi, politik, dan sosial.

#### 1.5. Sistematika Penulisan

Sistematika penulisan yang digunakan dalam Tugas Akhir ini terbagi dalam beberapa pokok bahasan, yaitu:

##### BAB I PENDAHULUAN

Bab pendahuluan membahas mengenai latar belakang, rumusan masalah, tujuan dan manfaat, serta ruang lingkup pelaksanaan Tugas Akhir “Klasifikasi Dokumen Berita Bahasa Indonesia Menggunakan Metode *Latent Dirichlet Allocation* (LDA) dan *Word2Vec*”.

##### BAB II TINJAUAN PUSTAKA

Bab ini membahas mengenai kajian pustaka yang berhubungan dengan Tugas Akhir sebagai landasan untuk merumuskan dan menganalisa permasalahan pada Tugas Akhir. Tinjauan pustaka yang digunakan meliputi klasifikasi teks, *preprocessing*, *Word2Vec*, *K-means clustering*, LDA, *Kullback Leibler Divergence*, dan evaluasi.

##### BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan mengenai langkah-langkah yang dilakukan dalam penelitian tugas akhir. Langkah-langkah tersebut diawali dengan gambaran umum penelitian, dilanjutkan dengan garis besar

penyelesaian masalah dalam bentuk blok proses. Langkah-langkah yang dilakukan meliputi pengumpulan data, *preprocessing*, *Word2Vec*, *K-means clustering*, *K-Fold*, pembentukan model, evaluasi, dan proses klasifikasi teks berita Bahasa Indonesia.

#### BAB IV HASIL DAN ANALISA

Bab ini menguraikan hasil skenario dan analisa eksperimen yang dimulai dari teknis pengumpulan data sampai hasil analisa dari setiap eksperimen yang dilakukan.

#### BAB V PENUTUP

Bab ini menjelaskan mengenai kesimpulan dari uraian yang telah dijabarkan pada bab-bab sebelumnya dan saran untuk pengembangan penelitian lebih lanjut.